Solution 1: Risk Minimizers for Generalized L2-Loss

(a) For the optimal constant model $f(\mathbf{x}) = \boldsymbol{\theta}$ for the loss $L(y, f(\mathbf{x})) = \big(m(y) - m(f(\mathbf{x}))\big)^2$, we first apply the following substitution $z^{(i)} = m(y^{(i)})$ for each $i = 1, \ldots, n$, and introduce $\boldsymbol{\theta}_m = m(\boldsymbol{\theta}) \in m(\mathbb{R})$. Note that the inverse of m is continuous and strictly monotone as well, so that the minimizer of the initial optimization problem, i.e.,

$$\arg\min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}} \sum_{i=1}^{n} (m(y^{(i)}) - m(\boldsymbol{\theta}))^2.$$

is the same as for the "substituted" optimization problem, i.e.,

$$m^{-1}\left(\arg\min_{\boldsymbol{\theta}_m \in m(\mathbb{R})} \sum_{i=1}^{n} (z^{(i)} - \boldsymbol{\theta}_m)^2\right).$$

For the term in the brackets we have seen in the lecture (optimizer of the empirical L2 risk) that

$$\arg\min_{\boldsymbol{\theta}_m \in m(\mathbb{R})} \sum_{i=1}^{n} (z^{(i)} - \boldsymbol{\theta}_m)^2 = \frac{1}{n} \sum_{i=1}^{n} z^{(i)} = \frac{1}{n} \sum_{i=1}^{n} m(y^{(i)}).$$

Consequently,

$$\hat{f}(\mathbf{x}) = m^{-1}\left(\frac{1}{n} \sum_{i=1}^{n} m(y^{(i)})\right)$$

is the optimal constant model for L.

(b) **Recommended solution:** We need to use the fact that $y, y^{(1)}, \ldots, y^{(n)}$ are i.i.d. and the arithmetic mean is an unbiased estimator. That being said,

$$\mathbb{E}_{xy}[m(y)] - \mathbb{E}_{xy}\left[\frac{1}{n} \sum_{i=1}^{n} m(y^{(i)})\right] = 0.$$

Therefore,

$$
\begin{aligned}
\mathcal{R}_L\left(\hat{f}\right) &= \mathbb{E}_{xy}[L\left(y, f(\mathbf{x})\right)] \\
&= \mathbb{E}_{xy}[\left(m(y) - m(\hat{f}(\mathbf{x}))\right)^2] \\
&= \mathbb{E}_{xy}\left[\left(m(y) - \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)^2\right] \\
&= \mathbb{E}_{xy}\left[\left(m(y) - \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)}) - \left(\mathbb{E}_{xy}[m(y)] - \mathbb{E}_{xy}\left[\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right]\right)\right)^2\right] \\
&= \mathbb{E}_{xy}\left[\left(m(y) - \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)}) - \mathbb{E}_{xy}\left[m(y) - \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right]\right)^2\right] \\
&= \mathsf{Var}\left(m(y) - \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right) \\
&= \mathsf{Var}(m(y)) + \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right) \\
&= \mathsf{Var}(m(y)) + \frac{1}{n^2}\mathsf{Var}(\sum_{i=1}^{n} m(y^{(i)})) \\
&= \mathsf{Var}(m(y)) + \frac{1}{n^2} \cdot n\mathsf{Var}(m(y^{(i)})) \\
&= \mathsf{Var}(m(y)) + \frac{1}{n^2} \cdot n\mathsf{Var}(m(y)) \\
&= \left(1 + \frac{1}{n}\right)\mathsf{Var}(m(y))
\end{aligned}
$$

Alternative solution:

Here we provide an alternative solution to (b). First, note that

$$
\begin{aligned}
\mathcal{R}_L\left(\hat{f}\right) &= \mathbb{E}_{xy}[L\left(y, f(\mathbf{x})\right)] \\
&= \mathbb{E}_{xy}[\left(m(y) - m(\hat{f}(\mathbf{x}))\right)^2] \\
&= \mathbb{E}_{xy}\left[\left(m(y) - \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)^2\right] \\
&= \mathbb{E}_{xy}\left[m(y)^2\right] - 2\mathbb{E}_{xy}\left[m(y)\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right] + \mathbb{E}_{xy}\left[\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\right].
\end{aligned}
$$

Now, because $y, y^{(1)}, \ldots, y^{(n)}$ are i.i.d. with $\mathbb{E}_{xy}\left[m(y^{(i)})\right] = \mathbb{E}_{xy}\left[m(y)\right]$, we get

$$
\begin{aligned}
\mathbb{E}_{xy}\left[m(y)\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right] &= \frac{1}{n}\mathbb{E}_{xy}\left[m(y)\sum_{i=1}^{n} m(y^{(i)})\right] \\
&= \frac{1}{n}\mathbb{E}_{xy}\left[m(y)\right]\mathbb{E}_{xy}\left[\sum_{i=1}^{n} m(y^{(i)})\right] \\
&= \frac{1}{n}\mathbb{E}_{xy}\left[m(y)\right] n\mathbb{E}_{xy}\left[m(y)\right] = \mathbb{E}_{xy}\left[m(y)\right]^2.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathbb{E}_{xy}\left[\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\right] &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}_{xy}\left[m(y^{(i)})\left(\sum_{i=1}^{n} m(y^{(i)})\right)\right]\right)\\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}_{xy}\left[m(y^{(i)})^2 + \sum_{j\neq i} m(y^{(i)})m(y^{(j)})\right]\right)\\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}_{xy}\left[m(y^{(i)})^2\right] + \sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}_{xy}\left[m(y^{(i)})m(y^{(j)})\right]\right)\\
&= \frac{1}{n^2}\left(n\mathbb{E}_{xy}\left[m(y)^2\right] + \sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}_{xy}\left[m(y^{(i)})\right]\mathbb{E}_{xy}\left[m(y^{(j)})\right]\right)\\
&= \frac{1}{n^2}\left(n\mathbb{E}_{xy}\left[m(y)^2\right] + n(n-1)\mathbb{E}_{xy}\left[m(y)\right]^2\right)\\
&= \frac{1}{n}\mathbb{E}_{xy}\left[m(y)^2\right] + (1-\frac{1}{n})\mathbb{E}_{xy}\left[m(y)\right]^2 .
\end{aligned}
$$

So, combining the three later math displays, we obtain

$$
\begin{aligned}
\mathcal{R}_L\left(\hat{f}\right) &= \mathbb{E}_{xy}\left[m(y)^2\right] - 2\mathbb{E}_{xy}\left[m(y)\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right] + \mathbb{E}_{xy}\left[\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\right]\\
&= \mathbb{E}_{xy}\left[m(y)^2\right] - 2\mathbb{E}_{xy}\left[m(y)\right]^2 + \frac{1}{n}\mathbb{E}_{xy}\left[m(y)^2\right] + (1-\frac{1}{n})\mathbb{E}_{xy}\left[m(y)\right]^2\\
&= \left(1+\frac{1}{n}\right)\left(\mathbb{E}_{xy}\left[m(y)^2\right] - \mathbb{E}_{xy}\left[m(y)\right]^2\right)\\
&= \left(1+\frac{1}{n}\right)\mathsf{Var}(m(y)).
\end{aligned}
$$

(c) In order to derive the risk minimizer, we consider the unrestricted hypothesis space $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}\}$. By the law of total expectation

$$
\begin{aligned}
\mathcal{R}_L(f) &= \mathbb{E}_{xy}\left[L(y, f(\mathbf{x}))\right]\\
&= \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[L(y, f(\mathbf{x}))\ \mid \mathbf{x}\right]\right]\\
&= \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[(m(y) - m(f(\mathbf{x})))^2\ |\mathbf{x}\right]\right].
\end{aligned}
$$

Since $\mathcal{H}$ is unrestricted we can choose f as we wish: At any point $\mathbf{x} = \mathbf{x}$ we can predict any value c we want. The best point-wise prediction is

$$
f^*(\mathbf{x}) = \text{argmin}_c \mathbb{E}_{y|x}\left[(m(y) - m(c))^2 \mid \mathbf{x}\right] \overset{(*)}{=} m^{-1}\left(\mathbb{E}_{y|x}\left[m(y) \mid \mathbf{x}\right]\right),
$$

where $(*)$ is due to

$$
\begin{aligned}
\text{argmin}_c \mathbb{E}\left[(m(y) - m(c))^2\right] &= \text{argmin}_c \underbrace{\mathbb{E}\left[(m(y) - m(c))^2\right] - (\mathbb{E}[m(y)] - m(c))^2}_{=\mathsf{Var}[m(y)-m(c)]=\mathsf{Var}[m(y)]} + (\mathbb{E}[m(y)] - m(c))^2\\
&= \text{argmin}_c \mathsf{Var}[m(y)] + (\mathbb{E}[m(y)] - m(c))^2 = m^{-1}\left(\mathbb{E}[m(y)]\right),
\end{aligned}
$$

because $\mathsf{Var}[m(y)]$ does not depend on c. Note that we could have used a similar substitution as in (a) here to derive f^*. Furthermore, if we use $m(x) = x$ such that the considered loss coincides with the L2 loss, we get (quite naturally) the same best point-wise prediction as for the L2 loss. Using an m corresponding to another notion of mean (e.g., harmonic or geometric mean), the best point-wise prediction for that other mean is obtained in each case.

(d) The optimal constant model in terms of the (theoretical) risk can be obtained from the previous by forgetting the conditioning on point $\mathbf{x} = \mathbf{x}$, which leads to

$$
\bar{f}(\mathbf{x}) = m^{-1}\left(\mathbb{E}_y\left[m(y)\right]\right).
$$

The risk of the latter is $\mathsf{Var}(m(y))$:

$$\mathcal{R}_L\left(\bar{f}\right) = \mathbb{E}_{xy}[\left(m(y) - m(\bar{f}(\mathbf{x}))\right)^2] = \mathbb{E}_y[\left(m(y) - \mathbb{E}_y\left[m(y)\right]\right)^2] = \mathsf{Var}(m(y)).$$

(e) The Bayes regret can be decomposed as follows:

$$\mathcal{R}_L\left(\hat{f}\right) - \mathcal{R}_L^* = \underbrace{\left[\mathcal{R}_L\left(\hat{f}\right) - \inf_{f\in\mathcal{H}} \mathcal{R}_L(f)\right]}_{\text{estimation error}} + \underbrace{\left[\inf_{f\in\mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^*\right]}_{\text{approximation error}}.$$

If we consider as the hypothesis space $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R} \mid f(\mathbf{x}) = \boldsymbol{\theta}\ \forall \mathbf{x} \in \mathcal{X}\}$, i.e., the set of constant models, then the estimation error is

$$\mathcal{R}_L\left(\hat{f}\right) - \inf_{f\in\mathcal{H}} \mathcal{R}_L(f) = \underbrace{\mathcal{R}_L\left(\hat{f}\right)}_{\overset{(b)}{=}(1+\frac{1}{n})\mathsf{Var}(m(y))} - \underbrace{\mathcal{R}_L(\bar{f})}_{\overset{(d)}{=}\mathsf{Var}(m(y))} = \left(1 + \frac{1}{n}\right)\mathsf{Var}(m(y)) - \mathsf{Var}(m(y)) = \frac{1}{n}\mathsf{Var}(m(y)),$$

while the approximation error is

$$\begin{aligned}
\inf_{f\in\mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^* &= \underbrace{\mathcal{R}_L(\bar{f})}_{\overset{(d)}{=}\mathsf{Var}(m(y))} - \mathcal{R}_L(f^*) \\
&= \mathsf{Var}(m(y)) - \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[(m(y) - m(f^*(\mathbf{x})))^2 \mid \mathbf{x}\ \right]\right] \\
&= \mathsf{Var}(m(y)) - \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[\left(m(y) - m(m^{-1}(\mathbb{E}_{y|x}\left[m(y) \mid \mathbf{x}\right]))\right)^2 \mid \mathbf{x}\ \right]\right] \\
&= \mathsf{Var}(m(y)) - \mathbb{E}_x\Big[\underbrace{\mathbb{E}_{y|x}\left[(m(y) - \mathbb{E}_{y|x}\left[m(y) \mid \mathbf{x}\right])^2 \mid \mathbf{x}\ \right]}_{=\mathsf{Var}[m(y)\ \mid\ \mathbf{x}\ ]}\Big] \\
&= \mathsf{Var}(m(y)) - \mathbb{E}_x\left[\mathsf{Var}\left[m(y) \mid \mathbf{x}\ \right]\right] \\
&= \mathsf{Var}\Big(\mathbb{E}_{y|x}\left[m(y) \mid \mathbf{x}\ \right]\Big).
\end{aligned}$$

Note that the larger the sample size n the lower the estimation error, while the approximation error remains constant.